Process Capability and Data Contamination

DOI: 10.12776/QIP.V21I3.910

Filip Tošenovský, Josef Tošenovský

Received: 20 April 2017 Accepted: 28 October 2017 Published: 30 November 2017

ABSTRACT

Purpose: The paper centres on process capability and its relation to data contamination. Process capability may be distorted due to imprecise data. The paper analyses to what extent capability changes reflect problems in data so that the changes can be attributed to data sampling rather than the true performance of the process. This is important because it is usually much simpler to increase the precision of data sampling than the process itself.

Methodology/Approach: The paper has two major parts. In part one, effect of data contamination on the observed process characteristic is analysed. The effect is analysed using data obtained from simulated random drawings and the chi-squared test. In the other part, reaction of capability to data contamination is observed. The capability is measured by a univariate capability index.

Findings: Regarding the sensitivity of the index to contamination, it is different depending on the capability before the contamination. This leads to conclusions about when the company using the index should focus more on the way the data is measured, and when it should focus more on improving the process in question. The analysis shows that if the company is used to high levels of capability and records its drop, it is worth analysing its measurement system first, as the index is at higher levels more sensitive to data contamination.

Research Limitation/implication: The study concerns a single univariate index, and the contamination is modelled with only several probability distributions.

Originality/Value of paper: The findings are not difficult to detect, but are not known in practice where companies do not realize that problems with their process capability may sometimes lie in the data they use and not in the process itself.

Category: Research paper

Keywords: capability index; data contamination; index sensitivity

1 INTRODUCTION

Semi-finished and finished products are often accompanied upon their distribution by a technical report that gives details about their observed technical features, so that customers are aware of how to handle them safely and effectively. This is a result of legislative requirements, which above all try to protect public health and ensure that products serve their original intended purpose, and customers' needs which form the demand and define the character of the supplied commodity. The technical features are usually quantifiable characteristics, which means that their levels are measured by a measurement system of the product suppliers or their business partners. This applies to most economic sectors, involving both light and heavy industries. The numerical characteristics are usually random variables because their level is affected by many factors and not all of them are under control during the production. In other words, certain noise is present. Because of the uncertainty, capability indices are frequently required by customers so that there is a general idea about where the levels of the features can be expected and with what probability. The indices are useful before the production becomes a mass production, which otherwise is informative enough regarding the product quality level fluctuations due to a long production sample history. However, they are also calculated once the production has its history and becomes mature because it is necessary to adjust the early-production rough values of the indices and make them more precise, and also because there must be a controlling mechanism ensuring that the production still runs its originally planned course and doesn't divert from it to an unintended territory. Once the indices are calculated, the production level they reflect is either accepted, or it must be altered if it is not satisfactory. Adjusting production is a huge task for any company, though, because of all the factors that enter the process. The factors involve composition and amount of materials, the number and character of production machinery, the character of operators, production environment encompassing air humidity and temperature and other production conditions. It is a complex and difficult task (Rzevski and Skobelev, 2014) but spending time and financial resources on production changes to bring its level higher might be completely unnecessary due to the possible fact that the indices do not reflect the true situation. One of the reasons that this case may happen is a potential contamination of the data the indices are calculated from, whatever the source of the contamination. If data distortion is present, it will most likely be a result of the imperfect measurement system used (Automotive Industry Action Group, 2010). The system is never perfect, strictly speaking. The question then is whether and how the indices are related to the data contamination. Is there any relation at all which would allow for the capability index used to signal that maybe the management of the company in question should first check the measurement system before rushing to change the production, because the system imperfections may have distorted the true value of the index? This paper tries to answer these questions.

The paper is divided into several sections. The following two sections describe the methodology used and the most commonly known univariate indices, and they also select the index which is in common use. Further, assumptions are introduced in the two sections, based on which the whole analysis of the problem will be performed. Section four analyses the selected index and the final section formulates essential findings and conclusions.

2 METHODOLOGY

The paper consists of two major ideological parts. In the first part, potential effect of data contamination on the *probability distribution* of the evaluated process characteristic is analysed. The contamination is represented by a random variable the distribution of which is not normal, but shares some properties with normal distribution, which are known to be its typical features. These features include symmetry of its distribution and its zero expected or average value. Also, as is the standard way of proceeding, the true uncontaminated process characteristic and the variable representing contamination are considered statistically independent (Greene, 2011), and the effect of contamination, or the distortion of the original variable by contamination, is mathematically expressed as an addition. It is assumed that the original uncontaminated variable is normally distributed, which is something that should be checked whenever capability indices are used, since most indices require this assumption for their proper use. The effect of contamination is analysed using sample data obtained from simulated random drawings, and the chi-squared statistical test is used to see if there is any distortion in the distribution of the originally normal process characteristic. In the next part, reaction of process capability to the extent of contamination is observed, as well, using differential calculus (Larson and Edwards, 2013). The *sensitivity analysis* allows one to make conclusions about when data contamination could be suspected if the capability changes unexpectedly. The capability is measured by a selected univariate capability index.

3 CAPABILITY INDICES AND ANALYSIS ASSUMPTIONS

Many capability indices have been developed over the years (Tošenovský, 2007), and we shall concentrate only on a particular one. Our attention will also be turned to the univariate case, which is the usual case in practice. Although multivariate indices exist, as well, it is more difficult for many companies to use and interpret them in practice and so the one-dimensional indices still dominate when it comes to their use.

Several generations of univariate indices were defined and analysed in the past, one following another in an effort to remove theoretical problems of their predecessors. The C_p index is one of the oldest, comparing the tolerance prescribed by customer, a measure of variation allowed for the observed process

characteristic, to the true variation of the characteristic based on its normal probability model. As is known, the index does not reflect in any way the potential diversion of the expected value of the characteristic from the customer-defined target value, which is something of primary interest to customers, and so the index fell out of favour in many industrial organizations. Its successor, the C_{pk} index, was a refinement in the sense that it recognized both not complying with the requirement that the characteristic be in average equal to the target value and the variation of the characteristic. On the other hand, one of the demerits of the index is that it can be misused – not being able to keep the target value can be substituted by a lower variation of the characteristic to the extent that the index will remain at an acceptably high level. Customers not analysing the genesis behind the calculation of the index will thus be fooled into thinking that everything is well with the company whose index they observe. The efforts made at removing this theoretical drawback resulted over time in a next-generation capability index – the C_{pm} index. This index still allows the aforementioned substitution effect but only to a limited extent. Consequently, the room to cheat customers via process capability evaluation shrank considerably. Last but not least, a further fine-tuning of C_{pm} brought yet another improvement in the form of the C_{pmk} index. The C_{pmk} shares the positive features of its predecessor, but is stricter in the sense that it punishes the process decentralization more severely. This index is frequently used, and we shall concentrate on its properties in the next section.

Let us now turn our attention to the assumptions under which the analysis will be carried out. First, as already suggested by the choice of the univariate index, the case of two-sided tolerance interval defined for the characteristic of interest will be scrutinized. Also, given the index, it is assumed that the quality characteristic, to be denoted X, is normally distributed: $X \sim N$. Further, it is assumed that the characteristic is contaminated by a variable Y which represents data contamination, mostly due to an imperfect measurement system. The consequence of this situation is the fact that whoever tries to calculate process capability with the selected index doesn't base its calculation on the true realizations of the variable X, but on the realizations of the random variable $Z = X + Y$. The usual *theoretical* model for the contamination variable is $Y \sim N(0, \sigma_Y^2)$ (Greene, 2011), but the exact normality will hardly ever, if ever at all, be the case. The zero expected value is usually assumed because very often there is a negligeable systematic shift in the measurement system due to the common practice that operators and other parts of the measurement system are varied during the application of the system and so any potential systematic shift is close to zero *in average*. What is often a natural and acceptable assumption, as well, is that the variables X, Y are independent. This condition implies that for the variance of Z, $var(Z)$, we have $var(Z) = var(X) + \sigma_Y^2 = \sigma_X^2 + \sigma_Y^2$, regardless of whether the contamination is normally distributed or not. To sum up, when it comes to the two major statistical characteristics of Z, its expected value and variance, we have:

$$E(Z) = E(X) + 0 = \mu, \tag{1}$$

$$var(Z) = \sigma_X^2 + \sigma_Y^2. \tag{2}$$

As is known from the statistical theory, should both X and Y be normally distributed, Z would also be normally distributed. However, as has been outlined, Y is usually normally distributed only approximately, which may result in Z not being normally distributed. If this is the case, a change in the distribution of Z might serve as a signal that data impurity is present.

4 ANALYSIS

In this section, we shall analyse the numerical behaviour of the selected capability index, the C_{pmk} index, which is defined as:

$$C_{pmk} = \min\left(\frac{USL-\mu}{3\sqrt{\sigma_X^2+(\mu-T)^2}}, \frac{\mu-LSL}{3\sqrt{\sigma_X^2+(\mu-T)^2}}\right). \tag{3}$$

This part follows up the previous section which suggested that if X was normally distributed, but Y was not, though it shared some features with normally distributed variables, then it might be the case that the variable Z would not be normally distributed. Such an event would be welcome because one could become suspicious that the data is contaminated, if the variable of interest X has always been quite precisely normally distributed, and all of a sudden it isn't. The analytical section of the paper scrutinizes whether this occurs or not, using techniques of simulation. Data is generated from the distribution $N(\mu, \sigma_X^2)$ for diverse enough values of the two parameters, specifically for $\mu = 1, 5, 10, 50$ and $\sigma_X^2 = 1, 5, 10, 50$. These figures represent, according to practical experience, small values 1 and 5, a medium value 10 and a larger value 50. Also, data from a distribution similar to the normal distribution were generated to represent realizations of the variable Y. To do so, the Student's t distribution with various degrees of freedom (diverse enough values k = 10 and k = 50 were selected) serves as a similar distribution. As is known, this distribution is symmetric around zero, the expected value of the variable with such a distribution (Forbes, et al., 2010). The symmetry envokes a similarity to the normal distribution, and the zero expected value is in line with the commonly accepted assumption that the data contamination due to measurement errors is zero in average. The data sample sizes generated for the various parameters of the two distributions were equal to 100. Once the data sets for both variables are generated, realizations of Z become available, and so it can be tested by standard statistical methods whether Z can be consider a normally distributed variable or not. Except for the t

distribution, other distributions similar to the normal distribution, at least regarding their symmetry, exist, as well. We considered specifically the lognormal distribution $lnN(\mu, \sigma^2)$ with parameters $\mu = 1$ or 3 and $\sigma^2 = 0.1$ or 0.2 (Forbes, et al., 2010). Such parameters ensure a high level of symmetry of the distribution. For the lognormal case, the normal parameters considered were $\mu =$ 10 or 50 and $\sigma_X^2 = 5$ or 50. The rest of the procedure was the same as in the case of the t distribution. In both cases, the chi-square test of normality at a 5% significance level was used.

Tab. 1 shows the results of the test with respect to the t-distrubion-based simulation of Y. The p-values of the test indicate that in all but three cases, denoted by " * ", the hypothesis of normality of Z is accepted. The p-values are in nearly all the cases greater than 0.05.

Table1 – Test of Normality of Z with t-distributed Contamination

μ	σ^2	k	*p-value*
1	1	10	0.59
1	1	50	0.46
1	1	200	0.75
1	5	10	0.12
1	5	50	0.05*
1	5	200	0.55
1	10	10	0.14
1	10	50	0.19
1	10	200	0.47
1	50	10	0.46
1	50	50	0.61
1	50	200	0.5
5	1	10	0.84
5	1	50	0.15
5	1	200	0.84
5	5	10	0.35
5	5	50	0.26
5	5	200	0.27
5	10	10	0.82
5	10	50	0.87
5	10	200	0.81

μ	σ^2	k	*p-value*
5	50	10	0.22
5	50	50	0.09
5	50	200	0.04*
10	1	10	0.88
10	1	50	0.23
10	1	200	0.29
10	5	10	0.22
10	5	50	0.85
10	5	200	0.08
10	10	10	0.82
10	10	50	0.74
10	10	200	0.3
10	50	10	0.5
10	50	50	0.9
10	50	200	0.9
50	1	10	0.65
50	1	50	0.85
50	1	200	0.28
50	5	10	0.22
50	5	50	0.32
50	5	200	0.33
50	10	10	0.21
50	10	50	0.55
50	10	200	0.02*
50	50	10	0.59
50	50	50	0.24
50	50	200	0.24

The following Tab. 2 shows the results of normality testing for the case when the contamination component is modelled with a lognormal distribution with the aforementioned parameters. It can be seen again that in all but one case the hypothesis of normality of Z is accepted. The results here are even more convincing.

Table 2 – Test of Normality with Lognormal Contamination

$(N)\ \mu$	$(N)\ \sigma^2$	$(LN)\ \mu$	$(LN)\ \sigma^2$	*p-value*
10	5	1	0.1	0.67
50	5	1	0.1	0.85
10	50	1	0.1	0.9
50	50	1	0.1	0.86
10	5	1	0.2	0.9
50	5	1	0.2	0.85
10	50	1	0.2	0.92
50	50	1	0.2	0.54
10	5	3	0.1	0.78
50	5	3	0.1	0.7
10	50	3	0.1	0.37
50	50	3	0.1	0.77
10	5	3	0.2	0.03*
50	5	3	0.2	0.14
10	50	3	0.2	0.86
50	50	3	0.2	0.09

We may conclude at this stage that as long as the contamination is not normally distributed, but its distribution has a symmetric shape with heavier or less heavy tails than the normal distribution, the data at hand – the realizations of the variable Z – will in most cases manifest themselves as if they were drawn from a normal distribution. This suggests that there is no way of recognizing the potential presence of data contamination in a majority of usual cases.

Since the problem of contamination cannot be recognized securely from the data, the question is whether there is a way of detecting it from the computed capability index itself, especially in cases when the index has changed from its long-term familiar level. To find this out, it is necessary to explore the mathematical behaviour of the index and its dependence on the extent to which the data contamination is present. This is going to be examined in this part of the paper. Before doing so, we shall make some assumption about the index, which nonetheless does not place any restrictions on the generality of the conclusions. First, it can hardly be expected that the process whose capability is calculated is perfectly centralized, so $\mu \neq T$ is expected with certainty. This is always the case in practice because whenever it seems that the process is centralized, it is centralized only seemingly due to the fact that all measurement systems are imperfect and cannot perform measurements with absolute precision. Further, as

long as the amount of decentralization towards USL or LSL is the same, the index will be the same, given (3). In this respect, the index behaves symmetrically, and so one can use either $(USL-\mu)/3\sqrt{\sigma_X^2+(\mu-T)^2}$ or $(\mu-LSL)/3\sqrt{\sigma_X^2+(\mu-T)^2}$ for the analysis. In other words, it is not going to make a difference if $USL-\mu$ or $\mu-LSL$ is considered for the analysis, when $USL-\mu=\mu-LSL$. We shall use the former case and assume $T<\mu<USL$. Next, although USL naturally affects the value of the index, we are more interested in general conclusions of the form „faster/slower drop/rise of the index" and not so much in its absolute change given by a *specific* number. This intention implies that the level of USL is not going to change such general conclusions and therefore can be set arbitrarily. To give an example of this statement, if one uses a level USL_1 and arrives at $C_{pmk}^{(1)}$, whereas someone else uses a level USL_2 and gets $C_{pmk}^{(2)}$, then $C_{pmk}^{(2)}=kC_{pmk}^{(1)}$, where $k=C_{pmk}^{(2)}/C_{pmk}^{(1)}$ is a positive constant. Now, if a change in $\sigma^2=\sigma_X^2$, say $\sigma_2^2-\sigma_1^2$, alters the index $C_{pmk}^{(1)}$ less than a change $\sigma_4^2-\sigma_3^2$, the same will be true about the change of $C_{pmk}^{(2)}$, because $\Delta C_{pmk}^{(2)}=k\Delta C_{pmk}^{(1)}$, and $k>0$ remains the same. To see the latter, compare the value of k, k_2, after the change of the variance,

$$k_2=\frac{(USL_2-\mu)/3\sqrt{\sigma_2^2+(\mu-T)^2}}{(USL_1-\mu)/3\sqrt{\sigma_2^2+(\mu-T)^2}}, \tag{4}$$

with its value k_1 before the change of the variance,

$$k_1=\frac{(USL_2-\mu)/3\sqrt{\sigma_1^2+(\mu-T)^2}}{(USL_1-\mu)/3\sqrt{\sigma_1^2+(\mu-T)^2}}. \tag{5}$$

We have $k_2/k_1=1$. Thus, the upper limit has no effect on the general conclusions. Finally, we are interested in the amount of process decentralization, among other things, i.e. in the difference $(\mu-T)^2$, not in T itself. In other words, the target value can also be set arbitrarily, as long as it stays in the middle of the tolerance interval, since it is the deviation from T, not T, that matters.

Let us now imagine that we are to evaluate the process by estimating the value of the expression $(USL-\mu)/3\sqrt{\sigma_X^2+(\mu-T)^2}$, but instead we estimate something else: the value $(USL-\mu)/3\sqrt{\sigma_X^2+\sigma_Y^2+(\mu-T)^2}$ because we work with realizations of the variable Z instead of X due to the presence of data contamination (see formulas (1) and (2)). The question is how the index reacts to a change in the amount of data contamination, the contamination being measured by σ_Y^2, given the current level of the process, defined by σ_X^2 and $(\mu-T)$. The answers are revealed by the concept of *rate of change.*

The rates of change of C_{pmk} with respect to σ_Y^2 are:

$$\frac{\partial C_{pmk}}{\partial \sigma_Y^2} = \frac{-(USL - \mu)}{6\sqrt{[\sigma_X^2 + \sigma_Y^2 + (\mu - T)^2]^3}}, \tag{6}$$

where USL and T are arbitrary constants and $T < \mu < USL$. Using the concept of marginal changes, the result implies that the index drops whenever data contamination enters the calculations, but the drop realizes to various extent depending on the original process capability before the drop due to contamination. Given μ, the drop due to contamination decreases when the original capability is lower in terms of a higher σ_X^2, and an analogous result applies for a given variability σ_X^2: the greater the amount of decentralization of the process, the smaller the drop of the index due to contamination. If it happens, however, that there is no contamination in the data, the question of what could have caused a drop in the index may still arise. Was it a greater process variability or greater decentralization? Of course, the answer to this question is fairly straightforward and is given by the formula defining the index. Based on the behaviour of the index implied by (6), it is now possible to make some conclusions regarding the use of the results just presented.

5 CONCLUSION

Let us make some final remarks about the consequencies implied by the behaviour of the C_{pmk}. Any chance of recognizing the situation, when the data is contaminated, by observing potential deviations in the probability distribution of the quality characteristic of interest from its normal distribution is next to zero. Data contamination, as we have seen, does not change the normal distribution in a majority of cases, if the stochastic behaviour of the contamination is modelled with the well-known symmetric or near-symmetric distributions – the t-distribution or the lognormal distribution. What the analysis does show, however, is that if the company using the index is used to high levels of process capability and records its drop, it is worth analysing its measurement system first before proceeding to dismantle the much more complicated production structure with all its production inputs and conditions. The index may have declined due to imprecise data instead, as at its higher levels it is more sensitive to data contamination. On the contrary, when the production capability tends to be low for a longer period of time, its further deterioration is most likely a result of production inefficiencies, not imprecise data. In that case, methods such as FMEA should be used. FMEA is good for "optimizing a product or process design from the perspective of potential failures" (Vykydal, et al., 2013). If no data contamination is found, and the index drops from a high level, it is more likely due to increased process variability, to which the index is more sensitive at its high levels, whereas a drop in the index from its lower levels may be due to increased decentralization rather than increased variability, provided the decentralization greater than one half can be expected. This is implied by the second power in the denominator of (3). The logic behind the discussion on

whether it is a greater decentralization or higher variability that worsened the process capability is such that when decentralization can be suspected of causing a drop in the index, it should be explored first, since if it is the case, it will usually be a result of a systematic problem in the process, which is much easier to detect and remove, as compared to detecting and lowering a process variability induced by a *myriad* of factors small in their effect.

Of course, when the measurement system requires an inspection, an upgrade of this management subsystem should be approached. By viewing this subsystem as a part of the ISO-based quality management system, one may, for instance, fine-tune it through role-play simulation techniques, which have the potential to better describe its functioning (Zgodavová, Kisela and Sutoová, 2016). By improving the measurement system, no more encumberance is put upon the entire system through an introduction of yet another variability. A reasonable approach may also include more sophisticated decision-making techniques, especially when several measurement systems are compared with respect to their quality, the comparison taking into account more than feature of theirs. Models, based on utility these measurement systems provide to their users, can then be analysed (see, Krajňák and Krzikallová, 2016, for the description of such models).

ACKNOWLEDGMENTS

This paper was prepared under the specific research projects No. SP2017/66 and SP2017/63 conducted at the Faculty of Metallurgy and Materials Engineering, VŠB-TU Ostrava, with a support of the Ministry of Education of the Czech Republic.

REFERENCES

Automotive Industry Action Group, 2010. *Measurement Systems Analysis*. 4th ed. Southfield: AIAG.

Forbes, C., Evans, M., Hastings, N. and Peacock, B., 2010. *Statistical Distributions*. 4th ed. Hoboken: John Wiley & Sons.

Greene, W.H., 2011. *Econometric Analysis*. 7th ed. Upper Saddle River: Pearson.

Krajňák, M. and Krzikallová, K., 2016. Application of Decision Analysis at Trading of Goods in the Selected European Union Member State. In: VŠB-Technical University of Ostrava, *Proceedings of the 3rd International Scientific Conference International Conference on European Integration 2016.* Ostrava, Czech Republic, 19-20 May. Ostrava: VŠB-Technical University of Ostrava.

Larson, R. and Edwards, B.H., 2013. *Calculus*. 10th ed. Boston: Brooks/Cole.

Rzevski, G. and Skobelev, P., 2014. *Managing Complexity*. Southampton: WIT Press.

Tošenovský, J., 2007. *Ekonomické a technologické hodnocení způsobilosti procesů*. DTO CZ: Dům techniky.

Vykydal, D., Plura, J., Halfarová, P., Fabík, R. and Klaput, P., 2013. Use of Quality Planning Methods in Optimizing Welding Wire Quality Characteristics. *Metalurgija*, 54(4), pp.529-532.

Zgodavová, K., Kisela, M. and Sutoová, A., 2016. Intelligent approaches for an organisation's management system change. *The TQM Journal*, 28(5), pp.760-773.

ABOUT AUTHORS

Ing. Filip Tošenovský, Ph.D., assistant professor at the Dept. of Quality Management, VŠB-Technical University of Ostrava, 17. listopadu 15/2172, Ostrava 708 33, Czech Republic, e-mail: filip.tosenovsky@vsb.cz.

Prof. RNDr. Josef Tošenovský, CSc., professor at the Dept. of Quality Management, VŠB-Technical University of Ostrava, 17. listopadu 15/2172, Ostrava 708 33, Czech Republic, e-mail: josef.tosenovsky@vsb.cz.